KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

The impact of order variance amplification/dampening on supply chain performance

Robert N. Boute, Stephen M. Disney, Marc R. Lambrecht and Benny van Houdt

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0603

The impact of order variance amplification/dampening on supply chain performance

Robert N. Boute[1], Stephen M. Disney[2], Marc R. Lambrecht[1] and Benny Van Houdt[3]

[1] Research Center for Operations Management, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium. E-mail: robert.boute@econ.kuleuven.be, marc.lambrecht@econ.kuleuven.be

[2] Logistics Systems Dynamics Group, Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK. E-mail: disneysm@cardiff.ac.uk

[3] Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerpen, Belgium. Email: benny.vanhoudt@ua.ac.be

ABSTRACT

We consider a two echelon supply chain where a single retailer holds an inventory of finished goods to satisfy an i.i.d. customer demand, and a single manufacturer produces the retailer's replenishment orders on a make-to-order basis. The objective of this paper is to analyse the impact of the retailer's replenishment policy (order variance amplification/dampening) on supply chain performance. Inventory control policies at the retailer often transmit customer demand variability to the manufacturer, sometimes even in an amplified form (known as the bullwhip effect), thereby imposing high inventory and capacity costs on the manufacturer. Reducing the retailer's order variability is desirable for the manufacturer, but inflates the retailer's inventory requirements. We consider two strategies with regard to the production capacity. In a flexible capacity strategy, the manufacturer invests in excess capacity to guarantee constant lead times in order to keep inventories low. The amount of investment depends on the retailer's order pattern. In an inflexible capacity strategy, the capacity is limited and independent of the retailer's replenishment decision. This results in stochastic lead times, thereby inflating the retailer's inventory requirements.

Keywords: *production and inventory control, supply chain management, bullwhip, queueing, capacity-inventory trade-off*

1. INTRODUCTION

This paper explores a coordinative supply chain approach to control the demand variability that is propagated to upstream stages through appropriate downstream inventory management. Lee et al. (1997) describe a problem frequently encountered in supply chains, called the bullwhip effect: demand variability increases as one moves up the supply chain. This *amplified* order variability can have large upstream cost repercussions. Balakrishnan et al. (2004) emphasize the opportunities to reduce supply chain costs by *dampening* order variability. However, despite the fact that the manufacturer benefits from smooth production, retailers, driven by the goal of reducing inventory costs, prefer to use replenishment policies that chase demand rather than dampen consumer demand variability. Dampening variability in orders may have a negative impact on the retailer's customer service due to inventory variance increases (Disney and Towill 2003). In this paper we analyse the impact of the retailer's inventory policy on total supply chain performance.

In the remainder of this section we describe our model and its assumptions. In the next section we discuss the scenarios of a flexible and an inflexible capacity strategy. Section 3 is devoted to the downstream inventory policy and its impact on order variance. In section 4 we consider the lead time distribution and the distribution of the net stock. Section 5 describes the cost analysis and section 6 concludes.

1.1. Model description

We consider a two echelon supply chain with a single retailer and a single manufacturer. Every period, the retailer observes customer demand. If there is enough on-hand inventory available, the demand is immediately satisfied. If not, the shortage is backlogged. To maintain an appropriate amount of on-hand inventory, the retailer places a replenishment order with the manufacturer at the end of every period.

The manufacturer does not hold a finished goods inventory but produces the retailer's orders on a make-to-order basis. The manufacturer's production system is characterized by a single server queueing model that sequentially processes the ordered units one by one on a first-come-first-served basis. When the production is busy, the orders join a queue of unprocessed orders. Once the complete replenishment order is produced, it replenishes the retailer's inventory. The time from the moment an order is placed to the moment that it replenishes the retailer's inventory, is the replenishment lead time, T_p. The production process at the manufacturer implies that the retailer's replenishment lead times are stochastic and correlated with the order quantity.

1.2. Assumptions

- The sequence of events in a period is as follows. First receive goods from the upstream partner, then observe and satisfy demand and finally place a replenishment order.
- Customer demand D is independently and identically distributed (i.i.d.) over time with an arbitrary, finite, discrete probability distribution function $f_D(\cdot)$.
- If the inventory on hand at the end of the period is positive ($NS_t > 0$), a holding cost C_h per unit is incurred to carry inventory to the next period. If the inventory on hand is negative ($NS_t < 0$), a backlog cost C_b per unit shortage is incurred.
- The production ("service") time M of a single unit is deterministic. To ensure stability (of the queue), we assume that the utilization of the production facility (average batch production time divided by average batch interarrival time) is strictly smaller than one.
- Define the capacity K as the number of units that can be produced in a period. The capacity investment cost function is given by $C(K) = C_0 + C_K \cdot K$, where C_0 represents the fixed capacity investment cost and C_K is a constant, marginal capacity investment cost. The fixed cost may represent all the planning, real estate, and other costs that are (nearly) independent of the size of the capacity. The marginal capacity investment cost is the additional cost to add one unit of capacity. When the installed capacity is insufficient, a unit can be produced in overtime capacity at extra cost C_P. We assume that $C_K < C_P$, otherwise it would never be optimal to invest in capacity.

In figure 1 we summarize the assumed cost structure.

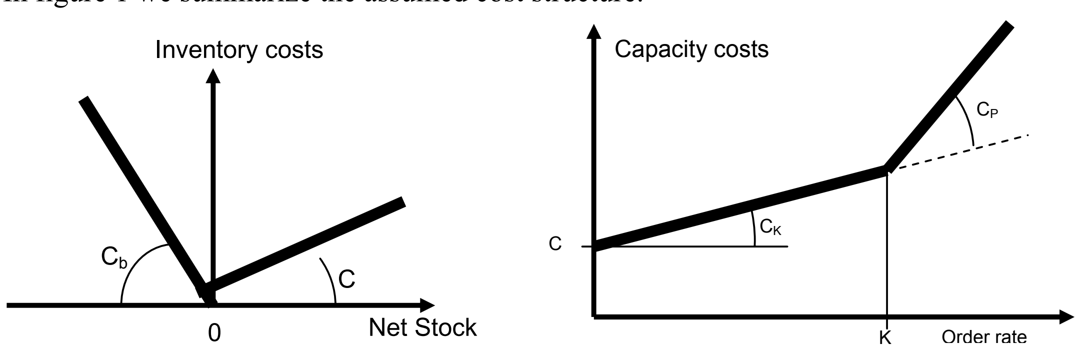

Figure 1: Cost structure of our research model

2. FLEXIBLE / INFLEXIBLE CAPACITY

The retailer's replenishment orders load the manufacturer's production system. We consider two strategies with regard to the production capacity. The first is a *flexible capacity* strategy. This means that the manufacturer invests in excess capacity in order to produce each order within the period after it was placed ($t_p = 0$). It is clear that when the orders fluctuate wildly, these investments will be larger than when the order pattern is less volatile. At the same time the inventory costs for the retailer are low since every order is replenished in the period after it was placed (zero lead times).

The second strategy is an *inflexible capacity* strategy, i.e., the manufacturer's capacity remains at a fixed level, irrespective of the retailer's order pattern. The manufacturer's capacity level may be lower than the maximum possible order quantity. As a result, when the available capacity in a period is insufficient to complete production of an order, then the next period's capacity is used to continue production of this order. The manufacturer delivers the retailer's orders as soon as the total order is produced, implying that lead times are variable and can be strictly positive. Moreover, when the retailer sends a volatile order pattern to the production queue, production (and delivery) lead times will be longer and more variable than when the retailer sends a constant order pattern to production. This in turn affects the retailer's inventory requirements.

In the remainder of this section we discuss the impact of both strategies on capacity investment and lead times.

2.1. Flexible Capacity – impact on capacity investment

Suppose the retailer wants the manufacturer to deliver the replenishment orders within the period after the order was placed (i.e., $t_p = 0$), then the production capacity has to be large enough to complete the production of each replenishment order within one time period. A key trade-off in capacity strategy is balancing the marginal cost of installed capacity C_K with the cost of capacity shortage (Van Mieghem 2006). In our case a capacity shortage implies a unit production in overtime capacity at cost C_P.

The installed capacity K is the number of units that can be produced in a period, and M is the production time of a single unit, expressed as a fraction of a period, or $K = M^{-1}$. The *capacity shortfall* in a given period measures how much of the period's order quantity exceeds available capacity, or equivalently, the number of units that are produced in overtime capacity in that period.

When the capacity is equal to the average order quantity, $K = E(O)$, the manufacturer experiences frequent capacity shortfalls if the order pattern is volatile. To counter the negative impact of volatility, it may be worth to invest in extra capacity above the average order quantity. The purpose of excess capacity is to provide a *safety capacity* to capture higher-than-expected orders. When the order volatility increases, the expected capacity shortfall will increase, but an investment in safety capacity can strongly reduce this capacity shortfall (Van Mieghem 2006).

A potential strategy is to set the capacity equal to the maximum order quantity, $K = O_{max}$, so that the capacity shortfall is zero and there is no production in overtime capacity. This would be a plausible strategy when the cost of production in overtime capacity is extremely large or when no overtime capacity is available. However, if for instance the order quantity reaches its maximum only occasionally, it may be beneficial to install a capacity $K < O_{max}$ and occasionally produce in overtime capacity at cost C_P (i.e., in case the order quantity exceeds the installed capacity). Hence the optimal capacity size K^* depends on the distribution of the replenishment orders placed by the retailer.

2.2. Inflexible Capacity – impact on lead times

The situation is totally different in the inflexible capacity strategy. When the order quantity exceeds the available capacity in a period, then the next period's capacity is used to continue the production of this order. Hence, there is no production in overtime capacity and the production of an order may be spread over several periods.

As the retailer's replenishment orders load the manufacturer's production, the nature of this loading process relative to the available capacity and the variability it creates determine the (production/replenishment) lead times. We actually extend a pure inventory system with *exogenous* lead times to a production-inventory system with *endogenous* lead times. The retailer's inventory replenishment lead times are "endogenously" determined by the manufacturer's production with limited capacity.

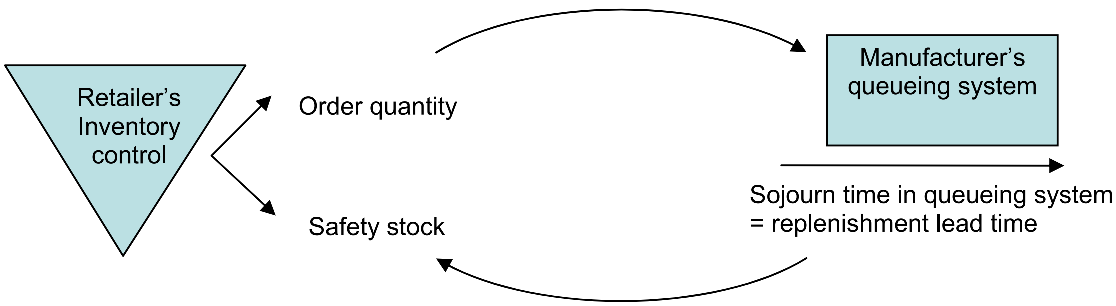

Figure 2: Interaction between retailer's inventory and manufacturer's production

In figure 2 the interaction between the retailer's replenishment policy and the manufacturer's production system is illustrated: the replenishment policy generates orders that constitute the arrival process at the manufacturer's queue. The time until the order is produced (the sojourn time in the queueing system), is the time to replenish the order. Hence, amplifying the order variability implies a more variable arrival pattern at the production queue, inducing longer and more variable lead times according to the laws of factory physics (Hopp and Spearman 2001). On the contrary, smoothing the order pattern results in shorter and less variable lead times. This replenishment lead time is a prime determinant in setting the safety stock requirements for the retailer.

3. DOWNSTREAM INVENTORY POLICY

3.1. Replenishment rule

Given the common practice in retailing to replenish inventories frequently (e.g. daily) and the tendency of manufacturers to produce to demand, we focus on periodic review, base-stock or order-up-to replenishment policies.

The *standard* periodic review base-stock replenishment policy is the (R,S) policy. At the end of every review period R, the retailer tracks his inventory position IP_t, which is the sum of the inventory on hand (that is, items immediately available to meet demand) and the inventory on order (that is, items ordered but not yet arrived due to the lead time) minus the backlog (that is, demand that could not be fulfilled and still has to be delivered). A replenishment order is then placed to raise the inventory position to an "order-up-to" or "base-stock" level S, which determines the retailer's order quantity in period t:

$$O_t = S - IP_t. \tag{1}$$

The base-stock level S is the inventory required to ensure a given customer service level. Orders are placed every R periods and after an order is placed, it takes T_p periods for the replenishment to arrive. Hence the risk period (the time between placing a replenishment order until receiving the subsequent replenishment order) is equal to the review period plus the replenishment lead time $R + T_p$. Since customer demand is i.i.d., the best estimate of all future demands is simply the long term average demand, E(D). Consequently, the base-stock level equals

$$S = [E(T_p) + R] \cdot E(D) + SS, \tag{2}$$

with SS denoting the retailer's safety stock.

In the remainder of this paper we assume that the review period R is one base period, i.e., we place an order at the end of every period. Substituting (2) into (1) we obtain

$$\begin{aligned} O_t &= E(D) + E(T_p) \cdot E(D) + SS - IP_t \\ &= E(D) + [DIP - IP_t], \end{aligned} \tag{3}$$

where $E(T_p) \cdot E(D) + SS$ can be seen as the *desired* inventory position DIP, which is the sum of the desired pipeline stock and desired net stock. The difference between the desired and actual inventory position $[DIP - IP_t]$ is denoted as the *inventory position deficit*.

Forrester (1961) and Magee (1958) introduce a proportional controller β into the inventory deficit, resulting in the following *generalised* order-up-to policy:

$$O_t = E(D) + \beta \cdot [DIP - IP_t], \tag{4}$$

with $0 < \beta < 2$. Forrester (1961) refers to $1/\beta$ as the "adjustment time". When $\beta < 1$ he explicitly acknowledges that the deficit recovery should be spread out over time, whereas $\beta > 1$ implies an overreaction to the inventory deficit. This replenishment rule is particularly powerful (Disney and Towill 2002) as it encompasses e.g. the way people play the Beer Game (Sterman 1989, Naim and Towill 1995), a general case of order-up-to policies and many variants of it (Dejonckheere et al. 2003), and with fine tuning it can reflect Materials Requirements Planning (Disney 2001). This "proportional order-up-to" policy is also equivalent to the "full-state order-to-up" policy, Gaalman and Disney (2006), assuming, as we do, an i.i.d. demand process.

3.2. Order variance amplification/dampening

When customer demand is i.i.d., the generalised replenishment policy generates an auto-correlated order pattern (see appendix A), given by

$$O_t = (1 - \beta) \cdot O_{t-1} + \beta \cdot D_t\,. \tag{5}$$

From this order "path" over time we can derive the steady state distribution of the order quantities given the finite, discrete demand distribution $f_D(\cdot)$. Let us denote the order distribution by $f_O(\cdot)$ and its corresponding cumulative order distribution by $F_O(\cdot)$.

Observe that when $\beta > 1$, the order pattern is negatively correlated and the generalised order-up-to policy may generate negative order quantities. Since in our model it is not possible to send negative orders to production, we have to preclude the possibility of negative orders. The following restriction on beta given the minimum and maximum demand ensures that $O_t \geq 1$ (see appendix B):

$$D_{min} + (1 - \beta) \cdot D_{max} \geq 2 - \beta. \tag{6}$$

To examine the variability in orders created by the generalised order-up-to policy, we look at the ratio of the variance of the orders over the variance of demand (in the literature this variance ratio is commonly used as a measure for the bullwhip effect). This can be easily derived from Eqn. (5):

$$\frac{\mathrm{Var}(O)}{\mathrm{Var}(D)} = \frac{\beta}{2-\beta}. \tag{7}$$

Hence, if we do not smooth, i.e. if $\beta = 1$, these expressions reduce to the standard base-stock policy, where $O_t = D_t$: we chase sales and thus there is no variance amplification. For $1 < \beta < 2$ we create bullwhip (variance amplification) and for $0 < \beta < 1$ we generate a smooth replenishment pattern (dampening order variability).

4. DETERMINATION OF LEAD TIMES AND INVENTORY

4.1. Determination of lead time distribution

The replenishment orders that load the production system are characterised by Eqn. (5). By analysing the characteristics of these replenishment orders, we implicitly analyse the characteristics of the production orders that arrive to the manufacturer's production system. As we can see from Eqn. (5), the generalised order-up-to policy generates batch arrivals with a fixed interarrival time (equal to the review period, $R = 1$) and with variable (autocorrelated) batch sizes.

In a separate paper (Boute et al. 2006) we develop a discrete time queueing model to estimate the lead time distribution based on *matrix analytic methods* (Neuts 1981, Latouche and Ramaswami 1999). In this paper, production times are phase type (PH) distributed, which can also be used to model deterministic production times. In addition we extend their model for values $\beta > 1$, taking restriction (6) into account. This queueing analysis returns the lead time distribution $f_{Tp}(\cdot)$ for each value of β.

Eqn. (7) shows that as β increases, the variability in the arrival pattern at the queue increases and consequently we expect longer and more variable lead times according to the laws of factory physics (Hopp and Spearman 2001). This intuition is confirmed for our specific queueing model (see Boute et al. 2006).

4.2. Determination of inventory distribution

When demand is probabilistic, there is a definite chance of not being able to satisfy some of the demand directly out of stock. Therefore, a buffer or safety stock is required to meet unexpected fluctuations in demand. We characterize the retailer's inventory random variable and use it to find its safety stock requirements. Due to the production process, lead times are stochastic and as a consequence we do not know exactly when a replenishment occurs.

We monitor the inventory on hand at the end of every period, after customer demand is observed and after a replenishment order has been placed. At the end of period t, there may be $k \geq 0$ orders waiting in the production queue and there is always 1 order in service (since the observation moment is immediately after an order placement) which is placed k periods ago (O_{t-k}). Note that k is a function of t, but we write k as opposed to k(t) to simplify the notation. In appendix C we show that the net stock distribution can then be written as

$$NS_t = SS - Z_t. \tag{8}$$

$$\text{with } Z_t = \sum_{i=0}^{k-1} D_{t-i} - E(T_p)\cdot E(D) + \sum_{i=k}^{t-1}(1-\beta)^{i-k}\cdot\left(D_{t-i} - E(D)\right). \tag{9}$$

The evolution of Z_t determines the evolution of the net stock NS_t. Since E(Z) = 0, E(NS)=SS. By means of the Markov process of the above mentioned queueing model, Boute et al. (2006) develop an algorithm to find the steady state distribution of Z_t, denoted by $f_Z(\cdot)$. The exact analysis is not straightforward due to the correlation between the different terms that make up Z_t. The value of D_{t-k} influences the age k of the current order in service: the larger the demand size, the larger the order size and consequently the longer it takes to produce the order. Moreover, since the order quantity is also affected by previously realised demand terms, the demand terms D_{t-i}, $i \geq k + 1$ also influence the order's age, k.

Given the distribution of Z, the amount of safety stock SS determines the corresponding inventory distribution $f_{NS}(\cdot)$. The value of SS is a decision variable and depends on the cost structure and the distribution of Z (see section 5). Since Z is function of β, SS is also impacted by the value of β.

In the flexible capacity scenario each replenishment order is produced within the period after it is placed, so that the production queue is always empty when an order is sent to production, or k = 0 in Eqn. (9). Moreover, since the lead time $t_p = 0$, Z_t simplifies to

$$Z_t = \sum_{i=0}^{t-1}(1-\beta)^i\left(D_{t-i} - E(D)\right), \tag{10}$$

and its steady state distribution $f_Z(\cdot)$ can be found from the compound demand distribution.

5. SUPPLY CHAIN PERFORMANCE

The objective of this paper is to measure the impact of the retailer's order decision (order variance amplification/dampening) on total supply chain performance. We consider the inventory costs at the retailer and the capacity costs at the manufacturer, and search for the value of the replenishment parameter β that minimises total supply chain costs for the flexible and inflexible capacity scenarios. Throughout this section we illustrate our analysis with a numerical example.

5.1. Numerical example

A retailer daily observes a customer demand which is randomly distributed between 21 and 40 units with an average of 30.5 units and a standard deviation of 7.5. The retailer replenishes his inventory with the generalised replenishment rule, i.e., he places orders at the end of every day equal to $O_t = E(D) + \beta \cdot [DIP - IP_t]$ (see Eqn. (4)).

When the replenishment parameter β < 1, the retailer sends a smooth, positively correlated order pattern to the manufacturer. When β > 1, the order pattern is negatively correlated with a larger variance than the observed customer demand. In order to exclude the possibility of negative order quantities, we limit the replenishment parameter to β < 1.525 (larger values of β may theoretically generate negative order quantities, see Eqn. (6)).

We assume the following cost components. A holding cost $C_h = 1$ is incurred per unit per day and a backlog cost $C_b = 20$ is incurred per unit that cannot be immediately satisfied from the inventory on hand. There is a fixed capacity investment cost $C_0 = 2$ and an additional cost per unit of installed capacity $C_K = 2$. A unit can be produced in overtime capacity at extra cost $C_P = 20$. Hence the capacity costs equal $C(K) = 2 + 2 \cdot K$.

5.2. Cost function

The capacity costs include the capacity investment cost C(K), which depends on the installed capacity K, and the number of units that are produced in overtime capacity in a period (in a flexible capacity strategy).

The inventory costs per period consist of a holding cost to keep a unit in inventory for a unit of time and a backlog cost for every unit of demand that can not be immediately fulfilled from the inventory on hand. Hence the inventory costs are equal to $C_h \cdot NS$ if $NS \geq 0$, and $C_b \cdot (-NS)$ if $NS < 0$. It is however more elegant to write the net stock NS as a function of the safety stock SS and the distribution of Z: $NS = SS - Z$.

The cost-minimisation problem can then be formulated as

$$\min_{\beta} \{ C_{INV}(SS^*, Z) + C_{CAP}(K^*, O) \}. \tag{11}$$

where

$$C_{INV}(SS^*, Z) = \min_{SS^*} \{ C_h \cdot E[(SS^* - Z)^+] + C_b \cdot E[(SS^* - Z)^-] \} \tag{12}$$

$$\begin{aligned} C_{CAP}(K^*, O) &= \min_{K^*} \{ C(K^*) + C_p \cdot E[(O - K^*)^+] \} && \text{when capacity is flexible,} \\ &= \min_{K^*} C(K^*) && \text{when capacity is inflexible.} \end{aligned} \tag{13}$$

The inventory and capacity costs depend on the amount of safety stock SS^*, the installed capacity K^* and the distribution functions of Z and O, which are function of the replenishment parameter β. In the remainder of this section we determine the safety stock SS^* and the installed capacity K^* that minimise respectively the inventory and capacity costs for a given value of β. In our analysis we make a distinction between the flexible and the inflexible capacity strategy.

5.3. Flexible Capacity Strategy

a) Optimal safety stock SS^* that minimises inventory costs for a given β

When capacity is flexible, the manufacturer invests in excess capacity in order to produce each order within the day after it is placed. Since lead times are zero in the flexible capacity strategy, we focus on the steady state distribution of Z_t given by Eqn. (10). A holding cost C_h is incurred when $NS > 0$, or equivalently, when $SS > Z$, and a backlog cost C_b is incurred when $NS < 0$, or $SS < Z$. The inventory cost function

$$C_{INV} = C_h \cdot E[(SS - Z)^+] + C_b \cdot E[(SS - Z)^-] \tag{14}$$

is then minimised by the critical fractile value, which provides the optimal stock out probability:

$$\Pr(NS < 0) = C_h / (C_h + C_b), \tag{15}$$

and the safety stock SS^* that minimises the inventory costs is then given by

$$\begin{aligned} \Pr(Z \leq SS^*) &= C_b / (C_h + C_b) \\ SS^* &= F_Z^{-1}(C_b / (C_h + C_b)), \end{aligned} \tag{16}$$

where $F_Z(\cdot)$ denotes the cumulative distribution function of Z. Substituting SS^* into Eqn. (14) provides the lowest inventory cost for a given value of β.

Clearly, as Z becomes more volatile, the optimal safety stock, SS^*, will increase, as well as the inventory costs. From Eqn. (10) we find that

$$\mathrm{Var}(Z) = \mathrm{Var}(D) \cdot 1 / \beta(2 - \beta) . \tag{17}$$

Hence, Z has a higher variance as we dampen the order pattern ($\beta < 1$) or as we amplify the orders ($\beta > 1$), compared to a pure chase sales policy ($\beta = 1$). As a result the inventory costs increase as we dampen or amplify the order variance, and are minimal when $\beta = 1$.

In figure 3 we plot the optimal safety stock SS^* of our numerical example that is required to maintain a 95.24% customer service level (according to Eqn. (15) the optimal stock-out probability is given by $C_h/(C_b+C_h) = 0.0476$)), together with its corresponding inventory costs C_{inv}. We observe that both the safety stocks and the inventory costs indeed increase as the order variance is dampened ($\beta < 1$) or amplified ($\beta > 1$), and the minimal inventory costs are found in a pure chase sales policy ($\beta = 1$). Overall, we observe that inventory costs are relatively low due to the zero lead times.

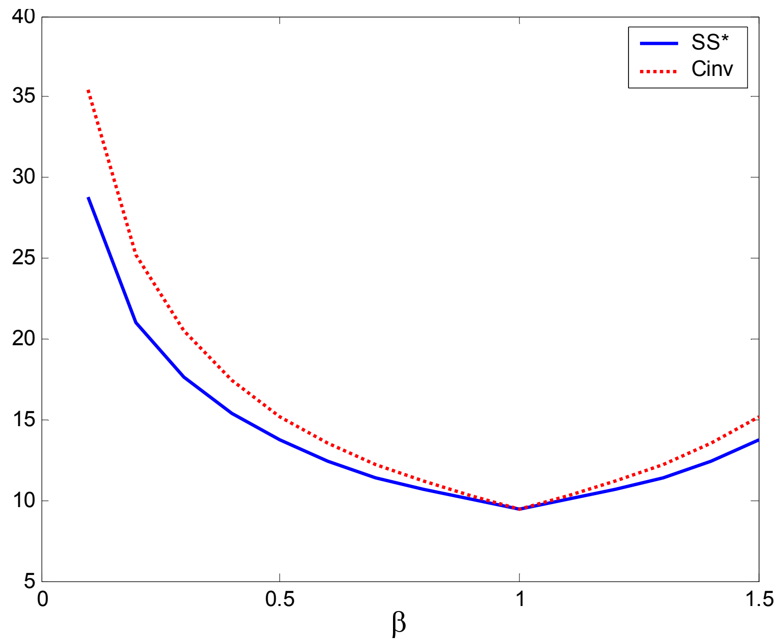

Figure 3 : Flexible capacity strategy: Impact of β on optimal safety stock SS^* and corresponding inventory costs C_{inv}

b) <u>Optimal capacity size K^* that minimises capacity costs for a given β</u>

In order to produce each order within one day time period (zero lead time), the manufacturer has to invest in capacity. The installed capacity K is the number of units that can be produced per period. The cost of this capacity investment is given by $C(K) = C_0 + C_K \cdot K$. When the order quantity exceeds the available capacity, then a cost C_P is incurred per unit that has to be produced in overtime capacity. The capacity cost function we have to minimise is consequently given by

$$C_{CAP} = C_0 + C_K \cdot K + C_P \cdot E[(O - K)^+]. \tag{18}$$

The optimal capacity size K^* that minimises this capacity cost function, satisfies a simple newsvendor solution. Van Mieghem (2006) shows that the optimal *capacity sizing condition* is given by:

$$\Pr(O > K^*) = C_K / C_P, \tag{19}$$

so that we find the optimal capacity size as

$$K^* = F_O^{-1}((C_P - C_K) / C_P), \tag{20}$$

with $F_O(\cdot)$ the cumulative order distribution function. It is intuitively clear that when the orders are less volatile, the optimal capacity size K^* will be lower than when orders are highly fluctuating. When the order pattern fluctuates wildly, it is preferable to invest in more capacity since production in overtime capacity is much more expensive. The optimal capacity size therefore depends on the retailer's ordering decision to amplify or dampen the order variance, or K^* is function of β.

In figure 4 we present the optimal capacity size K^* given the replenishment parameter β in our numerical example. As can be seen, K^* increases as the order pattern becomes more volatile (i.e., as β increases), and the corresponding capacity costs increase as well.

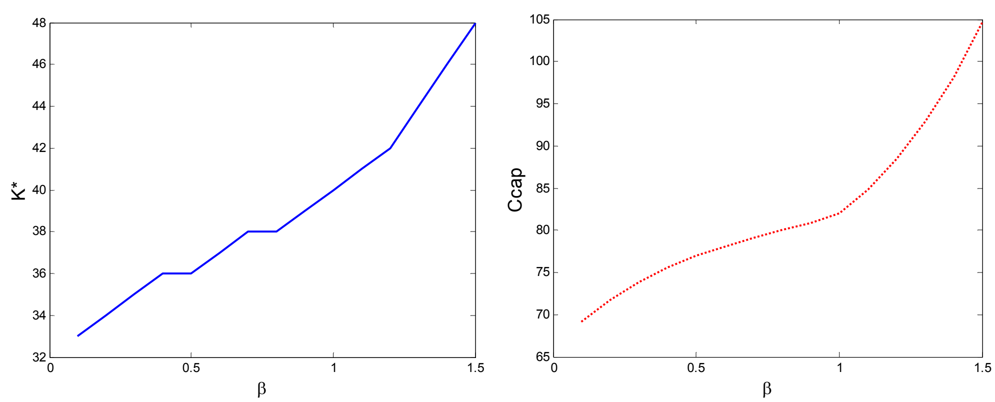

Figure 4: Flexible capacity strategy: Impact of β on optimal capacity size K^* and corresponding capacity costs C_{cap}

c) <u>Value of β that minimises total supply chain costs</u>

For a given value of the replenishment parameter β we described how to find the values of K^* and SS^* that minimise the capacity resp. inventory costs. The minimum total supply chain costs for a given value of β are found by simply adding up the minimum inventory and capacity costs corresponding to K^* and SS^*, since there is no interaction between inventory and capacity costs. We then seek the value of β that achieves the lowest total supply chain costs.

On the one hand, inventory costs show a U-shaped convex function of the parameter β with a minimum in β= 1. Order variance amplification or dampening both increase inventory costs compared to the chase sales policy. The capacity costs on the other hand increase as β increases. Compared to the chase sales policy, the capacity costs are lower when order variance is dampened and higher when the order variance is amplified. Hence, we may decrease total supply chain costs by dampening the orders to some degree. If we dampen the orders too much, capacity costs will further decrease, but the inventory costs may increase to a larger extent. The extent to which we should dampen the order variance depends on the relative costs of capacity and inventory.

In figure 5 we plot the total supply chain costs ($C_{INV} + C_{CAP}$) given the replenishment parameter β. Order variance amplification (β > 1) increases total supply chain costs sharply due to the reinforced effect of inventory and capacity costs. Order smoothing (β < 1) on the other hand exercises a dampening effect on total supply chain costs. Dependent on the relative size of capacity and inventory costs, order smoothing (to a certain extent) may decrease total supply chain costs. If we dampen the order variance to a great extent however (β < 0.5), total supply chain costs increase exponentially. In that case the decrease in capacity costs cannot compensate for the increase in inventory costs any more.

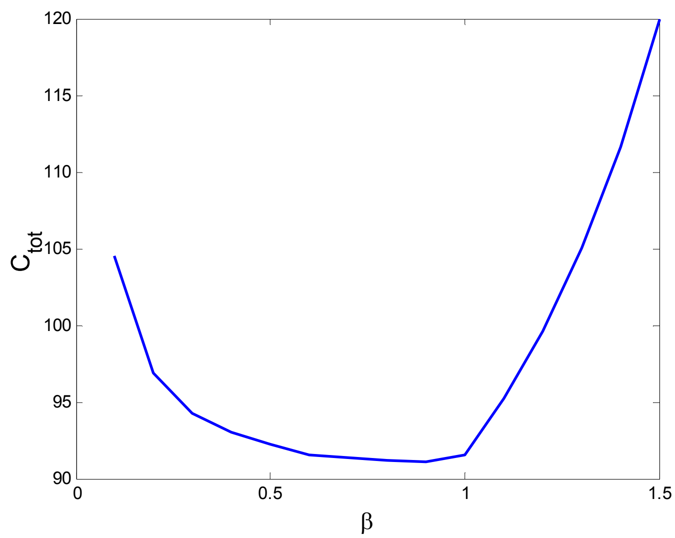

Figure 5: Flexible capacity strategy: Impact of the replenishment parameter β on total supply chain costs

5.4. Inflexible Capacity Strategy

a) Optimal safety stock SS^* that minimises inventory costs for a given β

In the inflexible capacity strategy the manufacturer's capacity remains at a fixed level, irrespective of the retailer's order pattern. Moreover, when the available capacity in a period is insufficient to complete the production of an order, then the next period's capacity is used to continue production of this order. We obtain a queueing model where lead times are stochastic and can be strictly positive, with increasing safety stocks and inventory costs as a consequence.

Similar to the flexible capacity strategy, the inventory cost function is given by

$$C_{INV} = C_h \cdot E[(SS - Z)^+] + C_b \cdot E[(SS - Z)^-]. \tag{21}$$

In this case however, Z is the steady state distribution of Z_t given by Eqn. (9). The distribution of Z is affected by β in two ways. First of all, similar to the flexible capacity strategy, the order variance has an impact on the variance of Z. Fluctuations are minimal in a pure chase policy (β = 1), and variability increases when orders are dampened (β < 1) or amplified (β > 1). But in the inflexible capacity strategy there is also a second factor that impacts the distribution of Z. The value of β also affects the lead time distribution. In section 4.1 we mentioned that lead times increase as β increases. This replenishment lead time is in turn a prime determinant of Z.

Hence we analyse the resulting impact of β on the distribution of Z. The safety stock SS^* that minimises the inventory costs is given by

$$SS^* = F_Z^{-1}(C_b / (C_h + C_b)). \tag{22}$$

Order variance dampening and amplification both increase the volatility of Z compared to a pure chase sales strategy, inflating the safety stock requirements. However, order variance dampening leads to lower and less variable lead times, exercising a compensating effect on the required safety stock. Order variance amplification on the other hand increases lead times, reinforcing the increased safety stock requirements.

b) <u>Optimal capacity size K^* that minimises capacity costs for a given β</u>

The capacity level remains fixed in the inflexible capacity strategy, independent of the order decision. Moreover there is no production in overtime capacity. The capacity cost function

$$C_{CAP} = C_0 + C_K \cdot K \tag{23}$$

is consequently minimised when the installed capacity K is as small as possible. The capacity K is however related to the utilization rate of the manufacturer's production system. The average utilization rate ρ is equal to the ratio of the average batch production time over the average inter-arrival time. Orders are placed every period, so the average inter-arrival time equals 1 period, and the average batch production time equals the average order quantity times the single unit production time: $\rho = E(O) \cdot M$. Hence

$$\begin{aligned} K &= M^{-1} \\ &= E(O) / \rho. \end{aligned} \tag{24}$$

Consequently, the capacity K^* has to be larger than the average order quantity E(O) in order to obtain a stable system with a production load smaller than one.

Suppose we assume a daily capacity equal to 32.5 units (at a fixed capacity cost of C_{CAP} = 67). This implies an average production load of $\rho = 30.5/32.5 = 0.9385$. The impact of β on the average lead time $E(T_p)$ and the optimal safety stock SS^* is shown in figure 6. As β increases, lead times increase as well since the order (arrival) pattern is more variable. This lead time effect has an impact on the optimal safety stock. We observe that the optimal safety stock indeed increases as the order variance is amplified ($\beta > 1$), but decreases when the order variance is dampened to some degree ($0.7 < \beta < 1$). The lead time reduction exercises a compensating effect on the optimal safety stock. However, when we dampen the order variance to a large extent ($\beta < 0.7$), the decrease in lead times cannot compensate the increase in inventory variability any more and safety stocks increase exponentially.

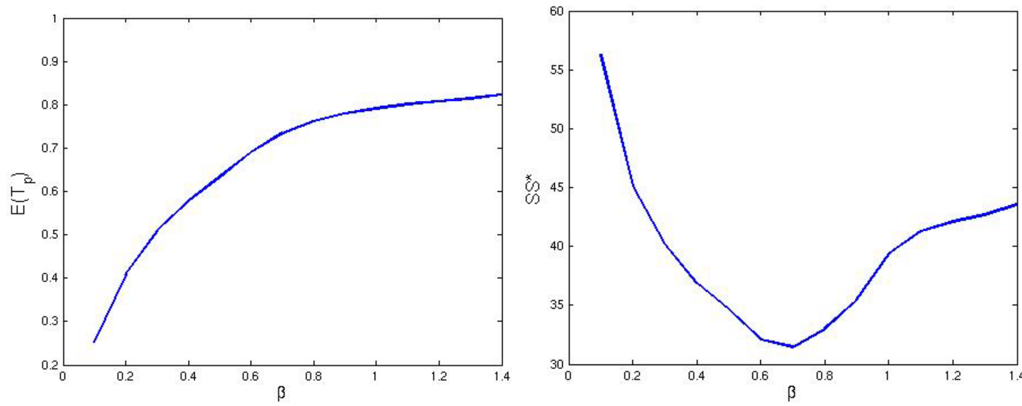

Figure 6: Inflexible capacity strategy: Impact of β on average lead time $E(T_p)$ and optimal safety stock SS*

c) <u>Value of β that minimises total supply chain costs</u>

For a given value of the replenishment parameter β we described how to find the value SS^* that minimises the inventory costs. The capacity costs are minimised when the installed capacity is as small as possible, provided that it exceeds the average order quantity. However, the amount of excess capacity has an impact on the lead times, which in turn determines the inventory costs. The larger the investment in capacity, the more the production load decreases, which in turn causes lead times to decrease. On the contrary, a minor investment in excess capacity corresponds to a high production load with long lead times as a result. These replenishment lead times determine the inventory costs at the retailer.

Consequently we need to trade-off capacity and waiting, which is in our case a *capacity-inventory trade-off*. For instance, as inventory costs are relatively less expensive, it is preferable not to invest in too much excess capacity. A high cost of inventory on the contrary increases the need for capacity investment.

In order to seek the lowest total supply chain costs, we assume a capacity size K and measure the impact of β on the inventory costs. Order variance amplification increases inventory variability ánd lead times, blowing up the inventory costs. Order variance dampening on the other hand gives rise to shorter and less variable lead times compared to the chase sales policy, which compensates the increase in inventory variability. The lowest inventory costs are found when we smooth the replenishment orders to some extent. If we smooth too much however, the lead time reduction cannot compensate the increase in inventory variability anymore and inventory costs increase.

To trade-off the cost of capacity against the cost of inventory, we change the capacity level K and measure its impact on inventory costs. It is clear that lead times (and inventory costs) decrease as capacity K increases (decreasing the utilization rate), but due to the complexity of our queueing model we cannot quantify the exact relation between the utilization rate and lead times analytically. Hence by means of a search procedure we determine the optimal capacity size K^* that minimises total supply chain costs. Obviously the value of K^* depends on the relative costs of capacity and inventory.

In figure 7 we show the impact of the parameter β on the inventory costs and on total supply chain performance (inventory and capacity costs).

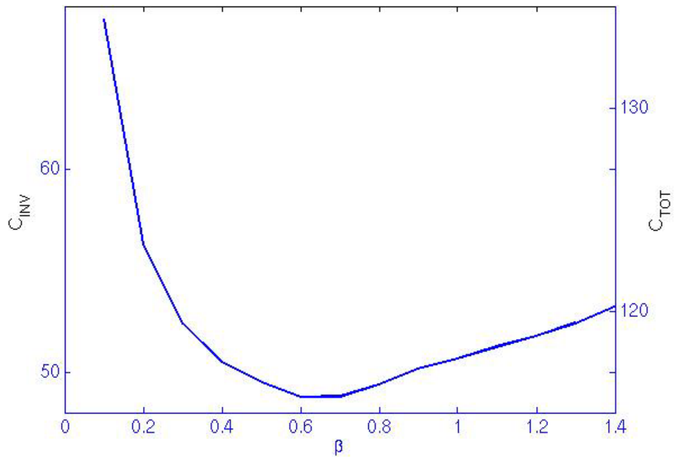

Figure 7: Inflexible capacity strategy: Impact of the replenishment parameter β on inventory costs and total supply chain costs when K = 32.5 (C_{CAP}=67)

Since capacity costs remain fixed, independent of β, total supply chain costs are only determined by the inventory costs. Analogous to the impact of β on the optimal safety stock, total supply chain costs increase when the order variance is amplified and decrease when the order variance is dampened to some degree ($0.6 < \beta < 1$). If we dampen the order variance to a large extent however ($\beta < 0.6$), total supply chain costs increase exponentially.

Suppose we increase the installed capacity slightly to K = 33 (at a total capacity cost of C_{CAP} = 68). This extra capacity investment decreases the average production load to $\rho = 30.5/33 = 0.9242$, which in turn causes lead times to decrease. Since lead times determine the optimal safety stocks, an investment in excess capacity will reduce the corresponding inventory costs.

In figure 8 we plot the impact of β on the resulting inventory costs and total supply chain costs when we increase the capacity to K = 33. The inventory costs are indeed significantly lower compared to a capacity K = 32.5. Moreover, we observe that the decrease in inventory costs compensates the increase in capacity costs since the total supply chain costs are lower then in figure 7. Hence it is beneficial to increase capacity (at extra cost) in order to improve total supply chain performance.

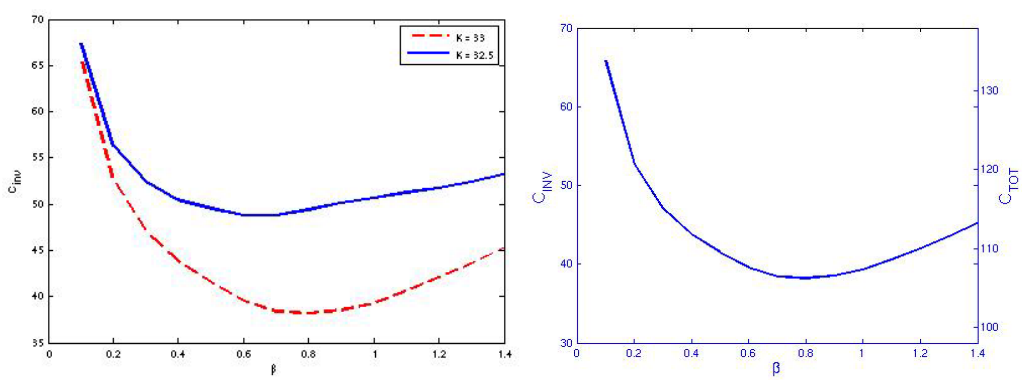

Figure 8: Inflexible capacity strategy: Impact of the replenishment parameter β on inventory costs and total supply chain costs when K = 33 (C_{CAP}=68)

5.5. Summary

This numerical example well illustrates the supply chain performance that results from the retailer's inventory decision and the manufacturer's strategy of a flexible or an inflexible capacity. Both in the flexible and inflexible capacity scenarios, order variance amplification increases total supply chain costs, and order variance dampening leads to lower supply chain costs. Consequently order smoothing is preferable. The degree to which we should smooth depends on the observed customer demand pattern and the cost components in the supply chain. In our example it is optimal to smooth the orders to a limited extent ($\beta \approx 0.7$).

The question whether companies should pursue a flexible or an inflexible capacity, depends on the relative size of capacity and inventory costs. The inflexible capacity strategy leads to higher inventory costs due to the stochastic lead times. These inventory costs may be reduced by increasing the installed capacity. Flexible capacity on the other hand requires production in overtime capacity and may require a large capacity investment, but has low inventory costs. In our numerical example total supply chain costs are lowest in a flexible capacity strategy.

6. CONCLUSIONS

In this paper we consider a downstream replenishment rule that is able to amplify or dampen the order variance upstream and we analyse the impact of this replenishment decision on total supply chain performance. The upstream echelon prefers a dampened or smooth order pattern from its downstream partner in order to minimise its own capacity costs. The latter however is not inclined to do so since a reduction in its order variance comes at the cost of an increased inventory cost. Both order variance amplification and dampening increase the inventory variability, inflating the safety stock requirements.

We propose a coordinative supply chain approach and consider two strategies investigating the capacity-lead time trade-off. When capacity is inflexible, order variance amplification results in longer and more variable lead times, increasing the inventory costs even more. At the same time order variance dampening decreases lead times, exercising a compensating effect on the retailer's inventories. When capacity is flexible, lead times can be reduced at the cost of extra capacity investment. However, as orders become more variable, the cost of extra capacity can be considerable. On the other hand, order smoothing decreases the capacity requirements. The lowest costs are in both scenarios found when orders are somewhat dampened. The extent to which we should dampen the order variance depends on the relative costs of capacity and inventory.

ACKNOWLEDGMENTS

This research contribution is supported by the contract grants G.0051.03 and G.0246.00 of the Research Programme of the Fund for Scientific Research — Flanders (Belgium) (F.W.O.-Vlaanderen). Benny Van Houdt is a postdoctoral Fellow of F.W.O.-Vlaanderen.

APPENDIX A: ORDER PATTERN GENERATED BY THE GENERALISED ORDER-UP-TO POLICY

In this appendix we show that the generalised order-up-to policy given by Eqn. (4) generates an auto-correlated order pattern given by

$$O_t = (1 - \beta) \cdot O_{t-1} + \beta \cdot D_t \,.$$

Proof. The generalised order-up-to policy generated orders according to

$$O_t = E(D) + \beta \cdot [DIP - IP_t].$$

Then,

$$\begin{aligned} O_t - O_{t-1} &= E(D) + \beta \cdot [DIP - IP_t] - E(D) - \beta \cdot [DIP - IP_{t-1}] \\ &= \beta \cdot (IP_{t-1} - IP_t). \end{aligned} \tag{A1}$$

The inventory position IP_t is monitored after customer demand is satisfied and before a replenishment order O_t is placed. Hence

$$\begin{aligned} IP_t &= IP_{t-1} + O_{t-1} - D_t \\ IP_{t-1} - IP_t &= D_t - O_{t-1} . \end{aligned} \tag{A2}$$

Substituting (A2) into (A1) results in

$$\begin{aligned} O_t - O_{t-1} &= \beta \cdot (D_t - O_{t-1}). \\ O_t &= (1 - \beta) \cdot O_{t-1} + \beta \cdot D_t . \end{aligned} \qquad \blacksquare$$

APPENDIX B: BOUNDS ON THE ORDER QUANTITIES GENERATED BY THE GENERALISED OUT POLICY

This section provides upper and lower bounds on the order quantities generated by the generalised order-up-to policy in Eqn. (4).
When $0 < \beta < 1$ the minimal and maximal order quantities are given by

$$\begin{aligned} O_{min} &= D_{min} \\ O_{max} &= D_{max}, \end{aligned}$$

since the generated order quantity is a simple exponential smoothing from the observed customer demand.
When $1 < \beta < 2$ we prove that the theoretical minimum and maximum order quantities are respectively given by

$$O_{min} = \frac{D_{min} + (1 - \beta) D_{max}}{2 - \beta}$$

$$O_{max} = \frac{D_{max} + (1 - \beta) D_{min}}{2 - \beta}.$$

Proof. Let the order quantity O_t reach its maximal value O_{max} in an arbitrary period t. Then, the order quantity in the next period t + 1 reaches its new minimum value O_{min} when the minimum demand realises, or

$$O_{t+1} = \beta \cdot D_{min} + (1 - \beta) \cdot O_t$$

$$= O_{min} \ .$$

Subsequently, a new maximum O_{max} is reached in the following period when the maximum demand is realised, or

$$\begin{aligned} O_{t+2} &= \beta \cdot D_{max} + (1 - \beta) \cdot O_{t+1} \\ &= O_{max} \ . \end{aligned}$$

Suppose the order pattern successively reaches its new minimum and maximum order quantity. Then, O_{2n} and O_{2n+1} are the respective minimum and maximum order quantities, given by

$$O_{min} = O_{2n} \quad = \beta \cdot D_{min} + (1 - \beta) \cdot O_{2n-1} \tag{A3}$$

$$O_{max} = O_{2n+1} = \beta \cdot D_{max} + (1 - \beta) \cdot O_{2n} \ . \tag{A4}$$

When $1 < \beta < 2$, we find that the minimum and maximum order quantities are respectively given by

$$O_{min} = \frac{D_{min} + (1-\beta)D_{max}}{2-\beta} \tag{A5}$$

$$O_{max} = \frac{D_{max} + (1-\beta)D_{min}}{2-\beta} \tag{A6}$$

Indeed, substituting (A5 – A6) into (A3 – A4) returns (A5 – A6) again. ■

Furthermore, using (A5), the restriction $O_{min} \geq 1$ can then be translated as

$$D_{min} + (1 - \beta) \cdot D_{max} \geq 2 - \beta \ .$$

APPENDIX C: DISTRIBUTION OF THE NET STOCK

In this section we derive an expression for the net stock distribution in function of the distribution of customer demand.

The inventory on hand NS_t at the end of period t is equal to the initial inventory on hand plus all replenishment orders received so far minus total observed customer demand. Since at the end of period t, the order O_{t-k} is in service, the orders placed more than k periods ago, i.e. O_{t-i}, $i \geq k+1$, are already delivered in inventory, while customer demand is satisfied up to the current period t. For our purposes the initial inventory level is a control variable, equal to the safety stock SS, determining the retailer's customer service. Since we assume that $O_t = D_t = E(D)$ for $t \leq 0$, the net stock after satisfying demand in period t is equal to

$$NS_t = SS + \left(E(T_p) + 1\right) \cdot E(D) + \sum_{i=k+1}^{t-1} O_{t-i} - \sum_{i=0}^{t-1} D_{t-i} \ . \tag{A7}$$

Substituting the auto-correlated order pattern (5) into (A7) gives

$$NS_t = SS + \left(E(T_p)+1\right)\cdot E(D) + \sum_{i=k+1}^{t-1}\left[(1-\beta)\cdot O_{t-i-1} + \beta\cdot D_{t-i} - D_{t-i}\right] - \sum_{i=0}^{k} D_{t-i}$$

$$= SS + \left(E(T_p)+1\right)\cdot E(D) + \sum_{i=k+1}^{t-1}\left[(1-\beta)\cdot O_{t-i-1} - (1-\beta)\cdot D_{t-i}\right] - \sum_{i=0}^{k} D_{t-i}.$$

Since $O_t = D_t = E(D)$ for $t \leq 0$, we find after backward substitution of Eqn. (5) that, for $t > 0$,

$$O_t = (1-\beta)^t \cdot E(D) + \sum_{j=1}^{t} \beta(1-\beta)^{j-1} D_{t-j+1},$$

so that we obtain

$$NS_t = SS + \left(E(T_p)+1\right)\cdot E(D) + \sum_{i=k+1}^{t-1}\left[(1-\beta)^{i-k}\cdot E(D) + \sum_{j=1}^{t-i-1}\beta(1-\beta)^{j} D_{t-i-j} - (1-\beta)\cdot D_{t-i}\right] - \sum_{i=0}^{k} D_{t-i}$$

$$= SS + E(T_p)\cdot E(D) + \sum_{i=k}^{t-1}\left[(1-\beta)^{i-k}\cdot\left(E(D) - D_{t-i}\right)\right] - \sum_{i=0}^{k-1} D_{t-i}.$$

REFERENCES

Balakrishnan, A. , Geunes, J. and Pangburn, M. (2004). Coordinating supply chains by controlling upstream variability propagation. *Manufacturing & Service Operations Management*, 6(2), pp 163-183.

Boute, R.N., Disney, S.M., Lambrecht, M.R. and Van Houdt, B. (2006). An integrated production and inventory model to dampen upstream demand variability in the supply chain. *European Journal of Operational Research*. To appear.

Dejonckheere, J., Disney, S.M., Lambrecht, M.R. and Towill, D.R. (2003). Measuring and avoiding the bullwhip effect: A control theoretic approach. *European Journal of Operational Research,* 147, pp 567-590.

Disney, S.M. (2001). *The production and inventory control problem in Vendor Managed Inventory supply chains*. PhD thesis. Cardiff University.

Disney, S.M. and Towill, D.R. (2002). A robust and stable analytical solution to the production and inventory control problem via a z-transform approach. *Proceedings of the 12th International Conference on Production Economics*, Igls, Austria, pp. 37-47.

Disney, S.M. and Towill, D.R. (2003). On the bullwhip and inventory variance produced by an ordering policy. *Omega*, 31, pp. 157-167.

Forrester, J. (1961). *Industrial Dynamics*. MIT Press, Cambridge MA.

Gaalman, G. and Disney, S.M., (2006), "On the bullwhip effect of Order-Up-To policies for ARMA(2,2) demand and arbitrary lead-times", *Proceedings of the 14th International Working Conference on Production Economics*, Innsbruck, Austria, February, Vol. 4, pp. 55-64.

Hopp, W.J. and Spearman, M.L. (2001). *Factory Physics.* 2nd edn. Irwin, McGraw-Hill.

Hosoda, T. (2005). The principles governing the dynamics of supply chains. PhD thesis. Cardiff University.

Latouche, G. and Ramaswami, V. (1999). Introduction to Matrix Analytic Methods and stochastic modeling. SIAM. Philadelphia.

Lee, H. L., Padmanabhan, V. and Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, 43(4), pp 546-558.

Magee, J. F. (1958). *Production Planning and Inventory Control*. McGraw-Hill. New York.

Nahmias, S. (1997). Production and Operation Analysis. 3rd edn. McGraw-Hill.

Naim, M.M. and Towill, D.R. (1995). What's in the pipeline? *Proceedings of the 2nd International Symposium on Logistics*, pp. 135-142.

Neuts, M. (1981). Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach. John Hopkins University Press.

Sterman, J.D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, 35 (3), pp. 321-339.

Van Mieghem, J. A. (2006). Operations Strategy, principles and practice. *Preprint, Working draft.*

Zipkin, P. H. (2000). *Foundations of Inventory Management*. McGraw-Hill. New York.